ANALYSIS OF K- MEANSCLUSTERING ON UNIFORM AND NON-UNIFORM DATA SET

[1]GAMINI SHARMA, [2]VEENA YADAV

[1,2]Computer Engineering
E-mail: [1]gamini782@gmail.com, [2]veena.yadav@poornima.org

Abstract— Cluster Analysis is a process of aggregating theobjects into various groups on the basis of their inter-cluster and intra-cluster similarities. But since we have large data objects with wide variety which are collected from wide variety of sources and perhaps include outliers as well, cluster formation still faces challenges over it. It faces many disputes such as a high dimension of the dataset, arbitrarystructure of clusters, scalability, domain knowledge and noisy data. Currently there are tremendous clustering algorithms to cleave data effectively had been proposed to address various existing challenges. The purpose of our paper is to analyze K-Means clustering algorithm on various data sets. In this paper, we have focused on analyzing the behavior of k-means clustering with uniform and non-uniform data sets.

Keywords— Euclidean,Cluster analysis,Clustering, K-means.

I. INTRODUCTION

Data Mining deals with the mining or extractionof knowledge from vast amount of data collected from various sources. The nature and behavior of any kind or type of data could be easily estimated by using data mining methods. Recently we have seen a tremendous expansion in the quantity of useful data being stored in the electronic format at hazardous rate. DBMS gave entry to the data saved but this is only small part of what could be gained from the data[4].

Clustering vast amount of data is a troublesome task since theobjective is to find a relevant partition in an unsupervised way (i.e.without any kind of prior knowledge). The clusters so formed should have more similarity within the objects belonging to the same cluster while minimum similaritybetween the objects of different clusters in order to maintain the high cluster cohesiveness.

Contradiction and affinityof the objects with other cluster objects are evaluated on the basis of their aspect values. A number of aspects such as partitioning methods, density-based methods, hierarchical methods and grid-based methods are used to partition and generate numerous clustering algorithms[1].

II. K-MEANS CLUSTERING

K-means algorithm is a partitioning based algorithm which tend to divide or partition the data into K clusters (C1, C2, C3 … CK), denoted by their cluster means generated considering their similarities with other data objects belonging to the same cluster. The mean associated with the cluster for every iteration evaluates the final cluster head so formed. Every forthcoming iteration reducesthe error function leading to estimated cluster heads. The complexity of this clustering algorithm performed with I iterations operated on N instances, each represented by A attributes, is: O(I*K*N*A). This linear complexity is perhaps the cause of popularity of the K-means algorithm[5].

A. History of the K-means Algorithm

The term "K-means" was first used by James MacQueen in 1967, although the concept arrived in 1957.Although it wasn't advertised until 1982 thiscertified algorithm was first suggestedby Stuart Lloyd in 1957 as a code for pulse-code modulation.This algorithm may result in local optima since it is critical to initial cluster heads so formed arbitrary. This partitioning method comprise of k partitions of the data, where each partition denotes a cluster with respective cluster heads such that $k \leq n$ (data objects) [2]. Categorizing the data into k clustersmust fulfill the following requirements:

i)Every cluster must comprise ofone or more than onedata objects

ii) Every object must belong to exactly one cluster.

B. K-means clustering Algorithm

The outline of K-means clustering algorithm:

Input: n objects (or data sets) and a number K(depicting the number of clusters to be formed from given data sets).

Algorithm:

Step1. Initialize K centre heads by arbitrary selecting K objects from the given data sets.

Step2. Assign each object to the cluster which has the minimum distance with the existing cluster heads.

Step3. When all of them are assigned then re-evaluate the position of the k centroid.

Step4. Iterate Steps 2 and 3 until the terminating criteria is attained.

The terminating criteria of the K-means Algorithm is given as follows:

1. No noticeable change in the objects of all the clusters.

2.Squared error is lower than some small threshold value αth.

The squared error SE is given by[3],

$$SE=\Sigma\Sigma||p-mi||$$

where mi is the mean of all the instances in cluster cj such that SE(j) < αth. Hence, it can be summarized that this algorithm achieves local optimum rather than global which is perhaps its limitation needs to be overcome. Apart from this, this algorithm is critical to noise and outliers which means that even if the object is quite far away from the cluster head it is still considered in that cluster which in turn distorts the cluster shape.

II. EXPERIMENTAL ANALYSIS

For experimental analysis we have considered the following experiments:

A. Uniform distribution

In this experiment, we took data sets of uniform density and distribution. We have chosen three datasets with their ranging sizes comprising of 20, 50 and 80 elements. When clustering algorithm was applied on these data sets it resulted in desired output. The output containing three clusters were formed accordingly. We also observed that the steps required in clustering the different data sets were increasing with the increase in size of data sets.Thus, k means clusters the given data sets uniformly and could be used for uniform data sets clustering by partitioning them according to the required value of K.

TABLE1: OBSERVATIONS WITH DIFFERENT DATA SETS

No of elements	Cluster centers	No of iterations
20	m1=76.9 m2=40.8 m3=11.0	6
50	m1=15.13 m2=47.55 m3=82.82	7
80	m1=15.72 m2=45.92 m3=81.16	9

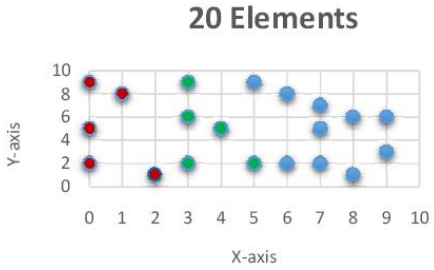

Fig. 1 Clustering of 20 elements.

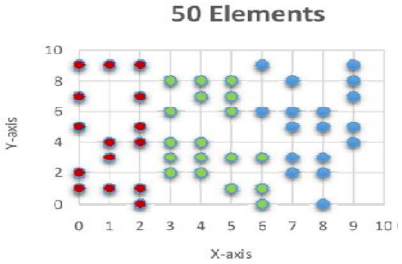

Fig. 2 Clustering of 50 elements

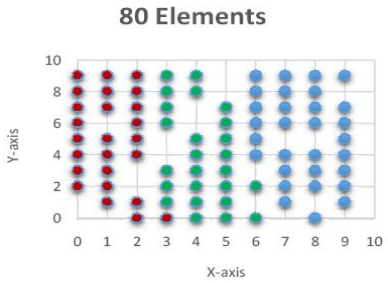

Fig. 3 Clustering of 80 elements

B. Non uniform distribution

In our second experiment, we considered data with different densities but uniform size. The data sets are chosen such that the distance between different objects is not uniform.

With different values of K chosen, k-means algorithm fails to build up clusters according to the requirements and thus could not be used for data sets of different densities or varying distances between data sets.

TABLE2: OBSERVATIONS WITH DATA SETS OF DIFFERENT DENSITIES

VALUE OF K	CLUSTER CENTRES
K=2	M1=76.842105 M2=24.0
K=3	M1=76.842105 M2=11.25 M3=31.2857142
K=4	M1=76.842105 M2=11.25 M3=31.285714 M4=NAN

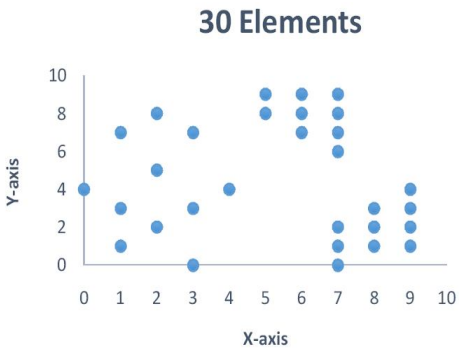

Fig. 4 Clustering of 30 elements

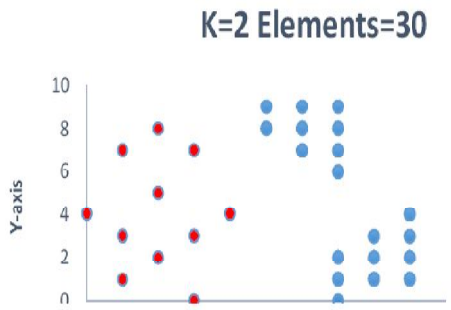

Fig. 5 Clustering of 30 elements

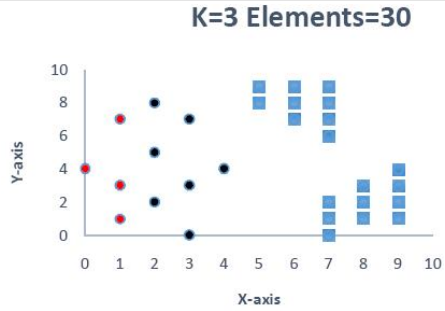

Fig. 6 Clustering of 30 elements

In our third experiment, we took non uniform data sets, that is, we took data sets with non-uniform distribution of values with varying sizes.

Table3: Observations of data sets of different sizes

VALUE OF K	CLUSTER CENTRES
K=2	M1=76.842105 M2=24.0
K=3	M1=76.842105 M2=11.25 M3=31.2857142
K=4	M1=76.842105 M2=11.25 M3=31.285714 M4=NAN

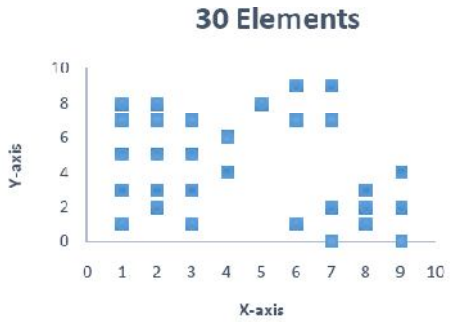

Fig7: Clustering of 30 elements

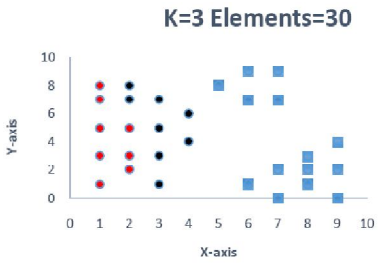

Fig.8 Clustering of 30 elements

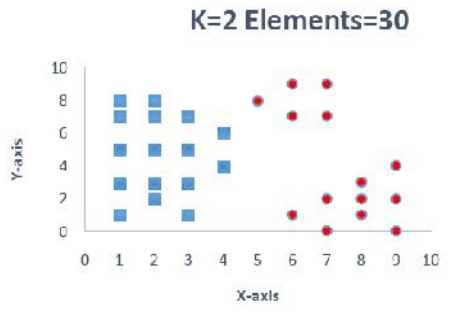

Fig.9 Clustering of 30 elements

As we could easily observe that the clusters formed were not found efficient or the data sets were misplaced in different clusters. When we considered the data sets with varied sizes, this algorithm failed to cluster the given data sets as required. Thus, this algorithm cannot be used with these type of data set values.

III. ADVANTAGES

The advantages of k-means clustering algorithm are as follows-

- If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls [6].
- K-Means produce comparatively tighter clusters in comparison to other clustering algorithms

IV. DISADVANTAGES

The disadvantages of k-means clustering algorithm as observed from analysis are as follows-

- It is difficult to predict the appropriate K value and is often critical.
- Different initial and arbitrary selected center heads can result in different final clusters formed.
- Its performance is critical with clusters (in the original data) of different size and different density distribution [8].

V. LIMITATIONS

- K-means is critical to clusters with different sizes, densities and non-globular shapes
- K-means has problems with data containing outliers or missing values.
- The number of clusters (K) required needs to be determined before running this algorithm. The algorithm is sensitive to an initial cluster head selection (starting cluster centroids) [7].
- The non-convex shape of clusters cannot be modeled by this algorithm.
- The number of iterations for bigger or huge data sets cannot be determined which is quite not relevant. It may take a huge number of iterations to converge thus affect the clustering of high dimensional data.

VI. FUTURE SCOPE

From the experimental result of K-means clustering algorithm results, the following considerations were made:

- There is no clustering algorithm that could be used to deal with wide variety and quantity of data produced from various sources. Usually, algorithms are designed with some assumptions and favor fewer type of biases.
- With this knowledge, we can say that, there is no bestclustering algorithm for all problems although some comparisons are possible. Thus, choosing a appropriate distance measure for k-means clustering algorithm can greatly lessen theburden of succeeding designs.

In this paper, we use prior knowledge of K-means clustering in the experimental result. With this knowledge, we can conclude that this algorithm is beneficial for practical evaluation in computing research areas.

CONCLUSION

Although K-means clustering cannot deal with non-uniform data sets but after cleaning the data values through various data mining methods, we could use this clustering algorithm which is efficient for uniform data sets with linear complexity. This could be used for spherical shaped clusters and also for others by generating variations in it. In this paper, we discussed the k-means algorithm which gives better efficiency but are more sensitive to noise and outliers.

REFERENCES

[1] ” A Comparative Study of Various Clustering Algorithms in DataMining,”(IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012..

[2] "Some Methods for classification and Analysis of Multivariate Observations". University of California Press. 1967, pp. 281–297.

[3] Lloyd, S. P. "Least square quantization in PCM". IEEE Transactions on Information Theory 28, 1982,pp. 129–137.

[4] “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, second Edition, (2006).

[5] “Cluster center initialization algorithm for kmeans clustering”, Pattern Recognition Letter.25, 2004, pp. 1293–1302.

[6] “The Global Kernel k-Means Algorithm for Clustering in Feature Space”, IEEE Trans. On Neural Networks, Vol. 20, No. 7, July 2009, pp. 1181-1194.

[7] “Survey of clustering algorithms”, IEEE Trans. Neural Networks., vol. 16, no. 3, 2005,pp. 645– 678.

[8] "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and SecurityInformatics (IITSI), pp.63-67, 2-4 April 2010.

★★★